A quick review:

After reading the paper by Keiser et al. (2007) [please give the full citation in Bibliography at the end of the summary.] describing the Similarity ensemble approach (SEA), which is used to predict binding based on chemical similarity of ligands, it appeared that in the process of creating the statistical model, the reasoning behind some of the steps was lacking or not provided.

In creating the statistical model, the following outline was followed:

1. Sets of ligands “across logarithmic set size intervals in the range of 10 to 1,000 molecules” were populated randomly.
2. For all pairs of ligands, one from set A the other from set B the Tanimoto coefficient score (Tc; ranging from 1, which means maximal similarity, to 0, which means minimal similarity) was calculated. The resulting sum of Tc scores (; raw score) was saved against the number of calculations made (*CS*; which is exactly the size of set A times the size of set B, which I’ll reference this as comparison size from now on).
3. Grouping by *CS* the mean and standard deviation (SD) of RS were calculated and a linear model was fit to *CS* against the mean while a non-linear model was fit for *CS* against SD.
4. For each comparison made normalized z-scores () were calculated as follows: .
5. To force the resulting distribution to match an extreme value distribution (EVD) a Tc cutoff of 0.57 was selected.

In particular three issues arose:

First, sampling was done “across logarithmic set size intervals in the range of 10 to 1,000 molecules” which itself is unclear and although the minimum of cs is 100 (between two groups of size 10) the model is later being used to conclude the significance of results even when the *CS* is less the a 100.

Second, SD was fitted using a non-linear regression and it is clear from the graph showing SD vs. *CS* that real SD values are getting further away from the predicted SD values as *CS* grows.

Third, it is not clear why the *ZS* distribution should be coerced into an EVD, the reasoning given in the original article cites blast as showing the EVD indicates biological significance, however in BLAST the underlying process is built in a way that maximizes the score (by finding the best alignment between two proteins).

Thus, we decided to derive a statistically robust model.

Methods:

All raw data used was provided by Keiser from a paper published in 2012 (Keiser et. Al 2012). This includes: A mapping between protein targets and their ligands, ligand fingerprints which were used to derive the statistical model and drug fingerprints which are not included in the ligand set, which were shown by (Keiser et. Al 2012) to bind to specific targets.

To try and solve the issues stated earlier we’ve repeated the original article’s sampling process in order to create a similar set of data going into the analysis step.

Sampling process:

1. Ligand sets of sizes ranging from 1 to 1,000 were randomly selected from a pool of 163,547 ligands.
2. For all pairs of ligands, L1 and L2, where one is from set A and the other from set B, the Tanimoto coefficient score (Tc; a measure of the chemical similarity between L1 and L2 which ranges between 0 for minimal similarity and 1 for maximal similarity) was calculated. The sum of Tc was saved, so a tuple was created : (, , ).

The sampling process was repeated 10 times.

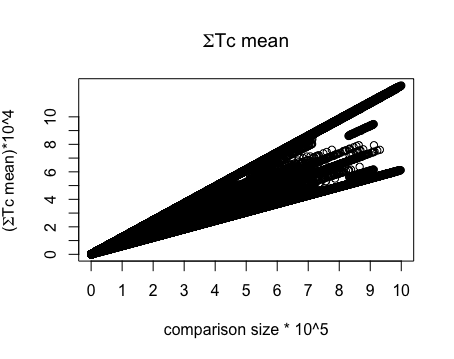
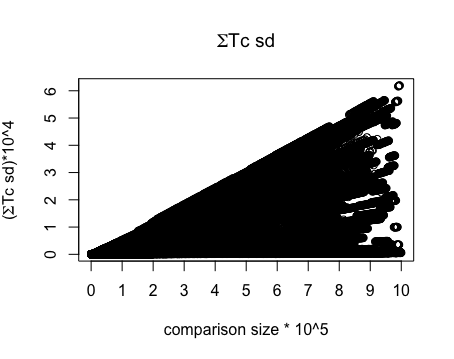
Analysis done:

In SEA (Keiser et al. 2007) sampling was done using “across logarithmic set size intervals in the range of 10 to 1,000 molecules” which resulted in set comparison sizes (the number of comparisons made between to sets, in the previous notation, ) ranging from 100 to . In that paper there were no sets of a single ligand in the sampling process. However, the same statistical model was also used in the next publication (Keiser et al. 2012) [add to bibliography] the same model is being used to predict the statistical significance of the score when a single ligand is checked against another set of ligands which sometimes resulted in a comparison size less than 100. Thus, we examined if the same results would be achieved when using set sizes ranging from 1 to 1,000 (using set size interval of 1).

Using the data sampled, a mapping between and was calculated in order to replicated the sampling data produced by Keiser et al. when creating the SEA statistical model.

The results are shown in Figure 1:

Figure 1

To the left is a plot showing the mean of for each comparison size (which equals in previous notation). When fitting a linear model to mean the model has an R-squared value of 0.8421.

To the right is a plot showing the standard deviation (SD) of for each comparison size.

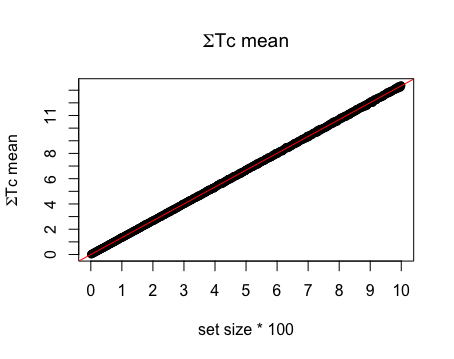
When fitting a linear model to SD the model has an R-squared value of 0.1325. It is easy to see that trying to fit any model to that data would not produce usable results. (In the paper describing SEA [Keiser 2007?] a nonlinear fit was computed). [The plot shows multiple linear lines. **Could you suggest why?** I’m not sure these are in fact multiple linear lines, I can try and create a model as I’ve done for the comparison of a single ligand and plot those lines on the two graphs.]

Next to tackle issues two and three we examined the possibility that the data above corresponds to a mixture. Specifically, that the results of comparisons of the same size may differ based on the sizes of the sets used. For example, a comparison of size 400 can be obtained by comparing one ligand to a set of 400 ligands, as well as by comparing two sets of 20 ligands each. We propose the idea that the results of the model are based on the size of the smaller of the two sets. To test this, the data was grouped based on the size of the smaller set of the two used for the comparison.

The results shown below are for a small set of one ligand only. As this is the most useful model.

The figure 2. shows the relation between the mean of to the size of the second set in the comparison (in the notation used before this plot shows , .

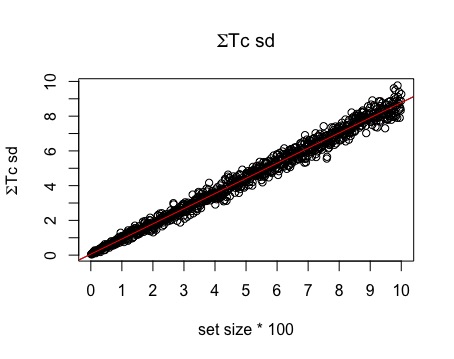
Figure 2



When fitting a linear model to mean the model has an R-squared value of 0.999.

The figure 3. shows the relation between the SD of to the size of the second set in the comparison (in the notation used before this plot shows , .

Figure 3



When fitting a linear model to SD the model has an R-squared value of 0.9868.

It seems that by taking into account the size of the smaller set we receive a better fit to the both the mean and SD and so now it is interesting to see how the distribution of *ZS* changes using the new predictors. In figure 4 we see a histogram of *ZS*. Looking at the histogram we were led to believe we are dealing with a bimodal normal distribution depicting two populations, the right most of ZS which belongs to ligands that will bind to tested targets and the left most with of ZS which belongs to ligands that will not bind to tested targets.

In order to test this, we’ve used drugs and targets that has been shown in-vitro by Keiser et al. (2012) to be positive predictions by the SEA model. Each drug’s *ZS* (for example Sertraline) was calculated against two target proteins groups, the first is human only, and the second is all organisms for which binding information is available. This is done since in Keiser et al. (2012) it is unclear whether the predictions were made against human proteins only or against all organisms for which binding information is available.

As you can see in figure 4. calculated *ZS* are plotted as vertical lines showing the resulting *ZS* these results strengthen the notion that we are dealing with an admixture of two populations.

It would be beneficial to contact Shoichet and Keiser and ask them to provide details of false positives and ambiguous results (predictions which do bind but at a high concentrations) and see if the by using our method the calculated *ZS* indeed belong the the second population or lay in between (around zero).

Figure 4

